# Discovering Latent Network Structure in Point Process Data

# Discovering Latent Network Structure in Point Process Data

**Scott W. Linderman** SLINDERMAN@SEAS.HARVARD.EDU
Harvard University, Cambridge, MA 02138 USA

**Ryan P. Adams** RPA@SEAS.HARVARD.EDU
Harvard University, Cambridge, MA 02138 USA

## Abstract

Networks play a central role in modern data analysis, enabling us to reason about systems by studying the relationships between their parts. Most often in network analysis, the edges are given. However, in many systems it is difficult or impossible to measure the network directly. Examples of latent networks include economic interactions linking financial instruments and patterns of reciprocity in gang violence. In these cases, we are limited to noisy observations of events associated with each node. To enable analysis of these implicit networks, we develop a probabilistic model that combines mutually-exciting point processes with random graph models. We show how the Poisson superposition principle enables an elegant auxiliary variable formulation and a fully-Bayesian, parallel inference algorithm. We evaluate this new model empirically on several datasets.

## 1. Introduction

Many types of modern data are characterized via relationships on a network. Social network analysis is the most commonly considered example, where the properties of individuals (vertices) can be inferred from "friendship" type connections (edges). Such analyses are also critical to understanding regulatory biological pathways, trade relationships between nations, and propagation of disease. The tasks associated with such data may be unsupervised (e.g., identifying low-dimensional representations of edges or vertices) or supervised (e.g., predicting unobserved links in the graph). Traditionally, network analysis has focused on *explicit network* problems in which the graph itself is considered to be the observed data. That is, the vertices are considered known and the data are the entries in the associated adjacency matrix. A rich literature has arisen in recent years for applying statistical machine learning models to this type of problem, e.g., Liben-Nowell & Kleinberg (2007); Hoff (2008); Goldenberg et al. (2010).

In this paper we are concerned with *implicit networks* that cannot be observed directly, but about which we wish to perform analysis. In an implicit network, the vertices or edges of the graph may not be directly observed, but the graph structure may be inferred from noisy emissions. These noisy observations are assumed to have been generated according to underlying dynamics that respect the latent network structure.

For example, trades on financial stock markets are executed thousands of times per second. Trades of one stock are likely to cause subsequent activity on stocks in related industries. How can we infer such interactions and disentangle them from market-wide fluctuations? Discovering latent structure underlying financial markets not only reveals interpretable patterns of interaction, but also provides insight into the stability of the market. In Section 4 we will analyze the stability of mutually-excitatory systems, and in Section 6 we will explore how stock similarity may be inferred from trading activity.

As another example, both the edges and vertices may be latent. In Section 7, we examine patterns of violence in Chicago, which can often be attributed to social structures in the form of gangs. We would expect that attacks from one gang onto another might induce cascades of violence, but the vertices (gang identity of both perpetrator and victim) are unobserved. As with the financial data, it should be possible to exploit dynamics to infer these social structures. In this case spatial information is available as well, which can help inform latent vertex identities.

In both of these examples, the noisy emissions have the form of events in time, or "spikes," and our intuition is that a spike at a vertex will induce activity at adjacent vertices. In this paper, we formalize this idea into a probabilis-

*Proceedings of the $31^{st}$ International Conference on Machine Learning*, Beijing, China, 2014. JMLR: W&CP volume 32. 

tic model based on mutually-interacting point processes. Specifically, we combine the Hawkes process (Hawkes, 1971) with recently developed exchangeable random graph priors. This combination allows us to reason about latent networks in terms of the way that they regulate interaction in the Hawkes process. Inference in the resulting model can be done with Markov chain Monte Carlo, and an elegant data augmentation scheme results in efficient parallelism.

## 2. Preliminaries

### 2.1. Poisson Processes

Point processes are fundamental statistical objects that yield random finite sets of events $\{s_n\}_{n=1}^N \subset \mathcal{S}$, where $\mathcal{S}$ is a compact subset of $\mathbb{R}^D$, for example, space or time. The Poisson process is the canonical example. It is governed by a nonnegative "rate" or "intensity" function, $\lambda(s) : \mathcal{S} \to \mathbb{R}_+$. The number of events in a subset $\mathcal{S}' \subset \mathcal{S}$ follows a Poisson distribution with mean $\int_{\mathcal{S}'} \lambda(s)\mathrm{d}s$. Moreover, the number of events in disjoint subsets are independent.

We use the notation $\{s_n\}_{n=1}^N \sim \mathcal{PP}(\lambda(s))$ to indicate that a set of events $\{s_n\}_{n=1}^N$ is drawn from a Poisson process with rate $\lambda(s)$. The likelihood is given by

$$p(\{s_n\}_{n=1}^N | \lambda(s)) = \exp\left\{-\int_{\mathcal{S}} \lambda(s)\mathrm{d}s\right\} \prod_{n=1}^N \lambda(s_n). \quad (1)$$

In this work we will make use of a special property of Poisson processes, the *Poisson superposition theorem*, which states that $\{s_n\} \sim \mathcal{PP}(\lambda_1(s) + \ldots + \lambda_K(s))$ can be decomposed into $K$ independent Poisson processes. Letting $z_n$ denote the origin of the $n$-th event, we perform the decomposition by independently sampling each $z_n$ from $\Pr(z_n = k) \propto \lambda_k(s_n)$, for $k \in \{1 \ldots K\}$ (Daley & Vere-Jones, 1988).

### 2.2. Hawkes Processes

Though Poisson processes have many nice properties, they cannot capture interactions between events. For this we turn to a more general model known as Hawkes processes (Hawkes, 1971). A Hawkes process consists of $K$ point processes and gives rise to sets of *marked* events $\{s_n, c_n\}_{n=1}^N$, where $c_n \in \{1, \ldots, K\}$ specifies the process on which the $n$-th event occurred. For now, we assume the events are points in time, i.e., $s_n \in [0, T]$. Each of the $K$ processes is a *conditionally Poisson process* with a rate $\lambda_k(t \,|\, \{s_n : s_n < t\})$ that depends on the history of events up to time $t$.

Hawkes processes have additive interactions. Each process has a "background rate" $\lambda_{0,k}(t)$, and each event $s_n$ on process $k$ adds a nonnegative impulse response $h_{k,k'}(t - s_n)$ to the intensity of other processes $k'$. Causality and locality of influence are enforced by requiring $h_{k,k'}(\Delta t)$ to be zero for $\Delta t \notin [0, \Delta t_{\mathsf{max}}]$.

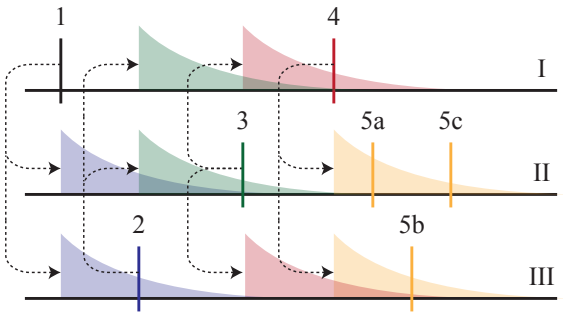


Figure 1: Illustration of a Hawkes process. Events induce impulse responses on connected processes and spawn "child" events. See the main text for a complete description.

By the superposition theorem for Poisson processes, these additive components can be considered independent processes, each giving rise to their own events. We augment our data with a latent random variable $z_n \in \{0, \ldots, n-1\}$ to indicate the cause of the $n$-th event (0 if the event is due to the background rate and $1 \ldots n-1$ if it was caused by a preceding event). The augmented Hawkes likelihood is then the product of likelihoods of each Poisson process:

$$p(\{(s_n, c_n, z_n)\}_{n=1}^N \,|\, \{\lambda_{0,k}(t)\}, \{\{h_{k,k'}(\Delta t)\}\}) = \prod_{k=1}^K p(\{s_n : c_n = k \wedge z_n = 0\} \,|\, \lambda_{0,k}(t)) \times \prod_{n=1}^N \prod_{k=1}^K p(\{s_{n'} : c_{n'} = k \wedge z_{n'} = n\} \,|\, h_{c_n,k}(t - s_n)),$$

where the densities in the product are given by Equation 1.

Figure 1 illustrates a causal cascades of events for a simple network of three processes (I-III). The first event is caused by the background rate ($z_1 = 0$), and it induces impulse responses on processes II and III. Event 2 is spawned by the impulse on the third process ($z_2 = 1$), and feeds back onto processes I and II. In some cases a single parent event induces multiple children, e.g., event 4 spawns events 5a-c. In this simple example, processes excite one another, but do not excite themselves. Next we will introduce more sophisticated models for such interaction networks.

### 2.3. Random Graph Models

Graphs of $K$ nodes correspond to $K \times K$ matrices. Unweighted graphs are binary adjacency matrices $\boldsymbol{A}$ where $A_{k,k'} = 1$ indicates a directed edge from node $k$ to node $k'$. Weighted directed graphs can be represented by a real matrix $\boldsymbol{W}$ whose entries indicate the weights of the edges. Random graph models reflect the probability of dif-

ferent network structures through distributions over these matrices.

Recently, many random graph models have been unified under an elegant theoretical framework due to Aldous and Hoover (Aldous, 1981; Hoover, 1979). See Lloyd et al. (2012) for an overview. Conceptually, the Aldous-Hoover representation characterizes the class of *exchangeable* random graphs, that is, graph models for which the joint probability is invariant under permutations of the node labels. Just as de Finetti's theorem equates exchangeable sequences to independent draws from a random probability measure , Aldous-Hoover renders the entries of $\boldsymbol{A}$ conditionally independent given latent variables of each node.

Empty graph models ($A_{k,k'} \equiv 0$) and complete models ($A_{k,k'} \equiv 1$) are trivial examples, but much more structure may be encoded. For example, consider a model in which nodes are endowed with a location in space, $\boldsymbol{x}_k \in \mathbb{R}^D$. This could be an abstract feature space or a real location like the center of a gang territory. The probability of connection between two notes decreases with distance between them as $A_{k,k'} \sim \text{Bern}(\rho e^{-||\boldsymbol{x}_k - \boldsymbol{x}_{k'}||/\tau})$, where $\rho$ is the overall sparsity and $\tau$ is the characteristic distance scale. This is known as a *latent distance model*.

Many models can be constructed in this manner. Stochastic block models, latent eigenmodels, and their nonparametric extensions all fall under this class (Lloyd et al., 2012). We will leverage the generality of the Aldous-Hoover formalism to build a flexible model and inference algorithm for Hawkes processes with structured interaction networks.

## 3. The Network Hawkes Model

In order to combine Hawkes processes and random network models, we decompose the Hawkes impulse response $h_{k,k'}(\Delta t)$ as follows:

$$h_{k,k'}(\Delta t) = A_{k,k'} W_{k,k'} g_{\theta_{k,k'}}(\Delta t). \tag{2}$$

Here, $\boldsymbol{A} \in \{0,1\}^{K \times K}$ is a binary adjacency matrix and $\boldsymbol{W} \in \mathbb{R}_+^{K \times K}$ is a non-negative weight matrix. Together these specify the *sparsity structure* and *strength* of the interaction network, respectively. The non-negative function $g_{\theta_{k,k'}}(\Delta t)$ captures the temporal aspect of the interaction. It is parameterized by $\theta_{k,k'}$ and satisfies two properties: a) it has bounded support for $\Delta t \in [0, \Delta t_{\mathsf{max}}]$, and b) it integrates to one. In other words, $g$ is a probability density with compact support.

Decomposing $h$ as in Equation 2 has many advantages. It allows us to express our separate beliefs about the sparsity structure of the interaction network and the strength of the interactions through a spike-and-slab prior on $\boldsymbol{A}$ and $\boldsymbol{W}$ (Mohamed et al., 2012). The empty graph model recovers independent background processes, and the complete graph recovers the standard Hawkes process. Making $g$ a probability density endows $\boldsymbol{W}$ with units of "expected number of events" and allows us to compare the relative strength of interactions. The form suggests an intuitive generative model: for each impulse response draw $m \sim \text{Poisson}(W_{k,k'})$ number of induced events and draw the $m$ child event times i.i.d. from $g$, enabling computationally tractable conjugate priors.

Intuitively, the background rates, $\lambda_{0,k}(t)$, explain events that cannot be attributed to preceding events. In the simplest case the background rate is constant. However, there are often fluctuations in overall intensity that are shared among the processes, and not reflective of process-to-process interaction, as we will see in the daily variations in trading volume on the S&P100 and the seasonal trends in homicide. To capture these shared background fluctuations, we use a sparse log Gaussian Cox process (Møller et al., 1998) to model the background rate:

$$\lambda_{0,k}(t) = \mu_k + \alpha_k \exp\{\boldsymbol{y}(t)\}, \quad \boldsymbol{y}(t) \sim \mathcal{GP}(\boldsymbol{0}, K(t,t')).$$

The kernel $K(t,t')$ describes the covariance structure of the background rate that is shared by all processes. For example, a periodic kernel may capture seasonal or daily fluctuations. The offset $\mu_k$ accounts for varying background intensities among processes, and the scaling factor $\alpha_k$ governs how sensitive process $k$ is to these background fluctuations (when $\alpha_k = 0$ we recover the constant background rate).

Finally, in some cases the process identities, $c_n$, must also be inferred. With gang incidents in Chicago we may have only a location, $\boldsymbol{x}_n \in \mathbb{R}^2$. In this case, we may place a spatial Gaussian mixture model over the $c_n$'s, as in Cho et al. (2013). Alternatively, we may be given the label of the community in which the incident occurred, but we suspect that interactions occur between clusters of communities. In this case we can use a simple clustering model or a nonparametric model like that of Blundell et al. (2012).

### 3.1. Inference with Gibbs Sampling

We present a Gibbs sampling procedure for inferring the model parameters, $\boldsymbol{W}$, $\boldsymbol{A}$, $\{\{\theta_{k,k'}\}\}$,$\{\lambda_{0,k}(t)\}$, and, if necessary, $\{c_n\}$. In order to simplify our Gibbs updates, we will also sample a set of parent assignments for each event $\{z_n\}$. Incorporating these parent variables enables conjugate prior distributions for $\boldsymbol{W}$, $\theta_{k,k'}$, and, in the case of constant background rates, $\lambda_{0,k}$. Detailed derivations are provided in the supplementary material.

**Sampling weights $\boldsymbol{W}$.** A gamma prior on the weights, $W_{k,k'} \sim \text{Gamma}(\alpha_W^0, \beta_W^0)$, results in the conditional dis-

tribution,

$$W_{k,k'} \,|\, \{s_n, c_n, z_n\}_{n=1}^N, \theta_{k,k'} \sim \mathrm{Gamma}(\alpha_{k,k'}, \beta_{k,k'}),$$
$$\alpha_{k,k'} = \alpha_W^0 + \sum_{n=1}^N \sum_{n'=1}^N \delta_{c_n,k} \delta_{c_{n'},k'} \delta_{z_{n'},n}$$
$$\beta_{k,k'} = \beta_W^0 + \sum_{n=1}^N \delta_{c_n,k}.$$

The posterior parameters correspond to the number of events caused by an interaction and the total unweighted rate induced by events on node $k$. Here and elsewhere, $\delta_{i,j}$ is the Kronecker delta function. We use the inverse-scale parameterization of the gamma distribution .

**Sampling background rates $\lambda_{0,k}$.** Similarly, for background rates $\lambda_{0,k}(t) \equiv \lambda_{0,k}$, the prior $\lambda_{0,k} \sim \mathrm{Gamma}(\alpha_\lambda^0, \beta_\lambda^0)$ is conjugate with the likelihood and yields the conditional distribution

$$\lambda_{0,k} \,|\, \{s_n, c_n, z_n\}_{n=1}^N, \sim \mathrm{Gamma}(\alpha_\lambda,\ \beta_\lambda),$$
$$\alpha_\lambda = \alpha_\lambda^0 + \sum_n \delta_{c_n,k} \delta_{z_n,0}$$
$$\beta_\lambda = \beta_\lambda^0 + T$$

This conjugacy no longer holds for log Gaussian Cox process background rates, but conditioned upon the parent variables, we must simply fit a log Gaussian Cox process for those events for which $z_n = 0$. We use elliptical slice sampling (Murray et al., 2010) for this purpose.

**Sampling impulse response parameters $\theta_{k,k'}$.** The logistic-normal density with parameters $\theta_{k,k'} = \{\mu, \tau\}$ provides a flexible model for the impulse response:

$$g_{k,k'}(\Delta t \,|\, \mu, \tau) = \frac{1}{Z} \exp\left\{ \frac{-\tau}{2} \left( \sigma^{-1}\left( \frac{\Delta t}{\Delta t_{\mathsf{max}}} \right) - \mu \right)^2 \right\}$$
$$\sigma^{-1}(x) = \ln(x/(1-x))$$
$$Z = \frac{\Delta t(\Delta t_{\mathsf{max}} - \Delta t)}{\Delta t_{\mathsf{max}}} \left( \frac{\tau}{2\pi} \right)^{-\frac{1}{2}}.$$

The normal-gamma prior $\mu, \tau \sim \mathcal{NG}(\mu, \tau | \mu_\mu^0, \kappa_\mu^0, \alpha_\tau^0, \beta_\tau^0)$ is conjugate and results in a conditional distribution with the following sufficient statistics:

$$x_{n,n'} = \ln(s_{n'} - s_n) - \ln(t_{\mathsf{max}} - (s_{n'} - s_n)),$$
$$m = \sum_{n=1}^N \sum_{n'=1}^N \delta_{c_n,k} \delta_{c_{n'},k'} \delta_{z_{n'},n},$$
$$\bar{x} = \frac{1}{m} \sum_{n=1}^N \sum_{n'=1}^N \delta_{c_n,k} \delta_{c_{n'},k'} \delta_{z_{n'},n} x_{n,n'}.$$

Intuitively, these correspond to the number of events caused by an interaction and their average delay.

**Collapsed Gibbs sampling $\boldsymbol{A}$ and $z_n$.** With Aldous-Hoover graph priors, the entries in the binary adjacency matrix $\boldsymbol{A}$ are conditionally independent given the parameters of the prior. The likelihood introduces dependencies between the rows of $\boldsymbol{A}$, but each column can be sampled in parallel. Gibbs updates are complicated by strong dependencies between the graph and the parent variables, $z_n$. Specifically, if $z_{n'} = n$, then we must have $A_{c_n,c_{n'}} = 1$. To improve the mixing of our sampling algorithm, first we update $\boldsymbol{A} \,|\, \{s_n, c_n\}, \boldsymbol{W}, \theta_{k,k'}$ by marginalizing the parent variables. The posterior is determined by the likelihood of the conditionally Poisson process $\lambda_{k'}(t \,|\, \{s_n : s_n < t\})$ (Equation 1) with and without interaction $A_{k,k'}$ and the prior comes from the Aldous-Hoover graph model. Then we update $z_n \,|\, \{s_n, c_n\}, \boldsymbol{A}, \boldsymbol{W}, \theta_{k,k'}$ by sampling from the discrete conditional distribution. Though there are $N$ parent variables, they are conditionally independent and may be sampled in parallel. We have implemented our inference algorithm on GPUs to capitalize on this parallelism.

**Sampling process identities $c_n$.** As with the adjacency matrix, we use a collapsed Gibbs sampler to marginalize out the parent variables when sampling the process identities. Unfortunately, the $c_n$'s are not conditionally independent and hence must be sampled sequentially.

**Computational concerns.** Compact impulse responses limit the number of potential event parents and significantly reduce the memory requirements and running time of our algorithm. If the average firing rate is constant, the expected number of potential parents per event will be linear in $K$. Summing the per-event contributions to the log likelihood can be done in $\mathrm{O}(\log N)$ time using standard parallel reductions. Hence, after parallelizing over the columns of $\boldsymbol{A}$ and the parents $z_n$, one step of our sampling algorithm takes $\mathrm{O}(K(K + \log N))$ time when process identities are known, and $\mathrm{O}((K+N)(K + \log N))$ time otherwise. On the datasets used in the following experiments, our GPU implementation[1] achieves 5-50 iterations per second.

## 4. Stability of Network Hawkes Processes

Due to their recurrent nature, Hawkes processes must be constrained to ensure their positive feedback does not lead to infinite numbers of events. A stable system must satisfy[2]

$$\lambda_{\mathsf{max}} = \max \,|\, \mathrm{eig}(\boldsymbol{A} \odot \boldsymbol{W}) \,| < 1$$

(c.f. Daley & Vere-Jones (1988)). When we are conditioning on finite datasets we do not have to worry about this. We simply place weak priors on the network parameters,

[1] https://github.com/slinderman

[2] In this context $\lambda_{\mathsf{max}}$ refers to an eigenvalue rather than a rate, and $\odot$ denotes the Hadamard product.

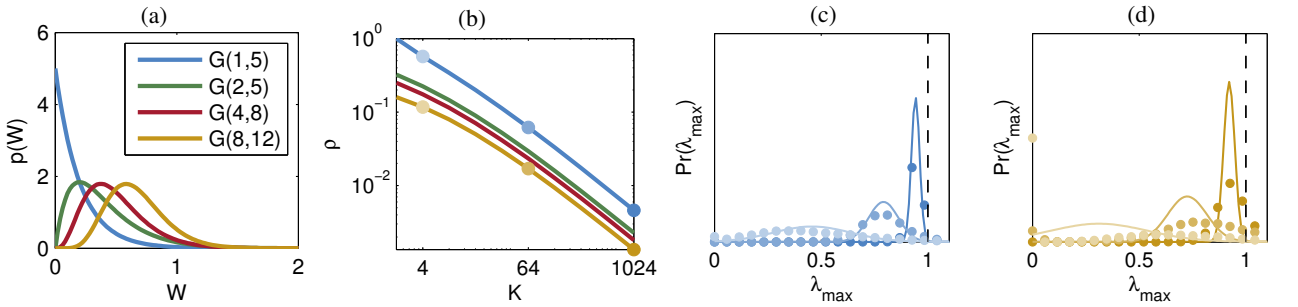


Figure 2: Empirical and theoretical distribution of the maximum eigenvalue for Erdős-Renyi graphs with gamma weights. (a) Four gamma weight distributions. The colors correspond to the curves in the remaining panels. (b) Sparsity that theoretically yields 99% probability of stability as a function of $p(W)$ and $K$. (c) and (d) Theoretical (solid) and empirical (dots) distribution of the maximum eigenvalue. Color corresponds to the weight distribution in (a) and intensity indicates $K$ and $\rho$ shown in (b).

e.g., a beta prior on the sparsity $\rho$ of an Erdős-Renyi graph, and a Jeffreys prior on the scale of the gamma weight distribution. For the generative model, however, we would like to set our hyperparameters such that the prior distribution places little mass on unstable networks. In order to do so, we use tools from random matrix theory.

The celebrated circular law describes the asymptotic eigenvalue distribution for $K \times K$ random matrices with entries that are i.i.d. with zero mean and variance $\sigma^2$. As $K$ grows, the eigenvalues are uniformly distributed over a disk in the complex plane centered at the origin and with radius $\sigma\sqrt{K}$. In our case, however, the mean of the entries, $\mu = \mathbb{E}[A_{k,k'}W_{k,k'}]$, is not zero. Silverstein (1994) has analyzed such "noncentral" random matrices and shown that the largest eigenvalue is asymptotically distributed as $\lambda_{\max} \sim \mathcal{N}(\mu K, \sigma^2)$.

In the simple case of $W_{k,k'} \sim \mathrm{Gamma}(\alpha, \beta)$ and $A_{k,k'} \sim \mathrm{Bern}(\rho)$, we have $\mu = \rho\alpha/\beta$ and $\sigma = \sqrt{\rho((1-\rho)\alpha^2 + \alpha)}/\beta$. For a given $K$, $\alpha$ and $\beta$, we can tune the sparsity parameter $\rho$ to achieve stability with high probability. We simply set $\rho$ such that the minimum of $\sigma\sqrt{K}$ and, say, $\mu K + 3\sigma$, equals one. Figures 2a and 2b show a variety of weight distributions and the maximum stable $\rho$. Increasing the network size, the mean, or the variance will require a concomitant increase in sparsity.

This approach relies on asymptotic eigenvalue distributions, and it is unclear how quickly the spectra of random matrices will converge to this distribution. To test this, we computed the empirical eigenvalue distribution for random matrices of various size, mean, and variance. We generated $10^4$ random matrices for each weight distribution in Figure 2a with sizes $K = 4$, 64, and 1024, and $\rho$ set to the theoretical maximum indicated by dots in Figure 2b. The theoretical and empirical distributions of the maximum eigenvalue are shown in Figures 2c and 2d. We find that for small mean and variance weights, for example $\mathrm{Gamma}(1, 5)$ in the Figure 2c, the empirical results closely match the theory. As the weights grow larger, as in $\mathrm{Gamma}(8, 12)$ in 2d, the empirical eigenvalue distributions have increased variance and lead to a greater than expected probability of unstable matrices for the range of network sizes tested here. We conclude that networks with strong weights should be counterbalanced by strong sparsity limits, or additional structure in the adjacency matrix that prohibits excitatory feedback loops.

## 5. Synthetic Results

Our inference algorithm is first tested on synthetic data generated from the network Hawkes model. We perform two tests: a) a link prediction task where the process identities are given and the goal is to simply infer whether or not an interaction exists, and b) an event prediction task where we measure the probability of held-out event sequences.

The network Hawkes model can be used for link prediction by considering the posterior probability of interactions $P(A_{k,k'} \mid \{s_n, c_n\})$. By thresholding at varying probabilities we compute a ROC curve. A standard Hawkes process assumes a complete set of interactions ($A_{k,k'} \equiv 1$), but we can similarly threshold its inferred weight matrix to perform link prediction.

Cross correlation provides a simple alternative measure of interaction. By summing the cross-correlation over offsets $\Delta t \in [0, \Delta t_{\max})$, we get a measure of directed interaction. A probabilistic alternative is offered by the generalized linear model for point processes (GLM), a popular model for spiking dynamics in computational neuroscience (Paninski, 2004). The GLM allows for constant background rates and both excitatory and inhibitory interactions. Impulse responses are modeled with linear basis functions. Area under the impulse response provides a measure of directed excitatory interaction that we use to compute a ROC curve. See the supplementary material for a detailed description of this model.

We sampled ten network Hawkes processes of 30 nodes each with Erdős-Renyi graph models, constant background rates, and the priors described in Section 3. The Hawkes processes were simulated for $T = 1000$ seconds. We used

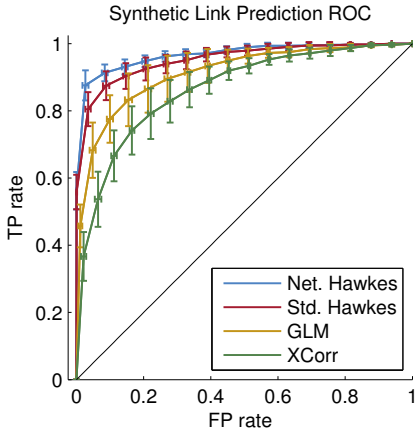


Figure 3: Comparison of models on a link prediction test averaged across ten randomly sampled synthetic networks of 30 nodes each. The network Hawkes model with the correct Erdős-Renyi graph prior outperforms a standard Hawkes model, GLM, and simple thresholding of the cross-correlation matrix.

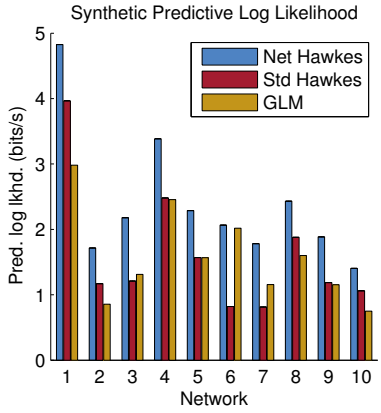


Figure 4: Comparison of predictive log likelihoods for the same set of networks as in Figure 3, compared to a baseline of a Poisson process with constant rate. Improvement in predictive likelihood over baseline is normalized by the number of events in the test data to obtain units of "bits per spike." The network Hawkes model outperforms the competitors in all sample networks.

the models above to predict the presence or absence of interactions. The results of this experiment are shown in the ROC curves of Figure 3. The network Hawkes model accurately identifies the sparse interactions, outperforming all other models. With the Hawkes process and the GLM we can evaluate the log likelihood of held-out test data. On this task, the network Hawkes outperforms the competitors for all networks. On average, the network Hawkes model achieves $2.2 \pm .1$ bits/spike improvement in predictive log likelihood over a homogeneous Poisson process. Figure 4 shows that on average the standard Hawkes and the GLM provide only 60% and 72%, respectively, of this predictive power. See the supplementary material for further analysis.

| Financial Model | Pred. log lkhd. (bits/spike) |
|---|---|
| Indep. LGCP | 0.594 |
| Std. Hawkes | 0.912 |
| Net. Hawkes (Erdős-Renyi) | 0.903 |
| Net. Hawkes (Latent Distance) | 0.888 |

Figure 5: Comparison of financial models on a event prediction task, relative to a homogeneous Poisson process baseline.

## 6. Trades on the S&P 100

As an example of how Hawkes processes may discover interpretable latent structure in real-world data, we study the trades on the S&P 100 index collected at 1s intervals during the week of Sep. 28 through Oct. 2, 2009. Every time a stock price changes by $\pm 0.1\%$ of its current price an event is logged on the stock's process, yielding a total of $K = 100$ processes and $N$=182,037 events.

Trading volume varies substantially over the course of the day, with peaks at the opening and closing of the market. This daily variation is incorporated into the background rate via a log Gaussian Cox process (LGCP) with a periodic kernel (see supplementary material). We look for short-term interactions on top of this background rate with time scales of $\Delta t_{\max} = 60$s. In Figure 5 we compare the predictive performance of independent LGCPs, a standard Hawkes process with LGCP background rates, and the net-

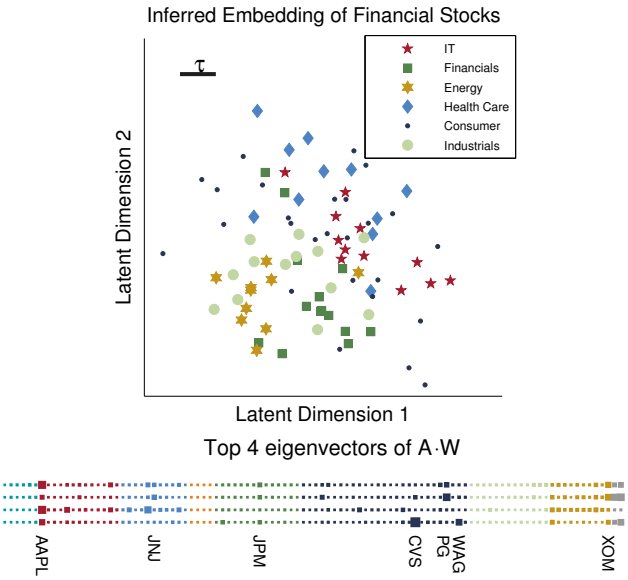


Figure 6: Top: A sample from the posterior distribution over embeddings of stocks from the six largest sectors of the S&P100 under a latent distance graph model with two latent dimensions. Scale bar: the characteristic length scale of the latent distance model. The latent embedding tends to embed stocks such that they are nearby to, and hence more likely to interact with, others in their sector. Bottom: Hinton diagram of the top 4 eigenvectors. Size indicates magnitude of each stock's component in the eigenvector and colors denote sectors as in the top panel, with the addition of Materials (aqua), Utilities (orange), and Telecomm (gray). We show the eigenvectors corresponding to the four largest eigenvalues $\lambda_{\max} = 0.74$ (top row) to $\lambda_4 = 0.34$ (bottom row).

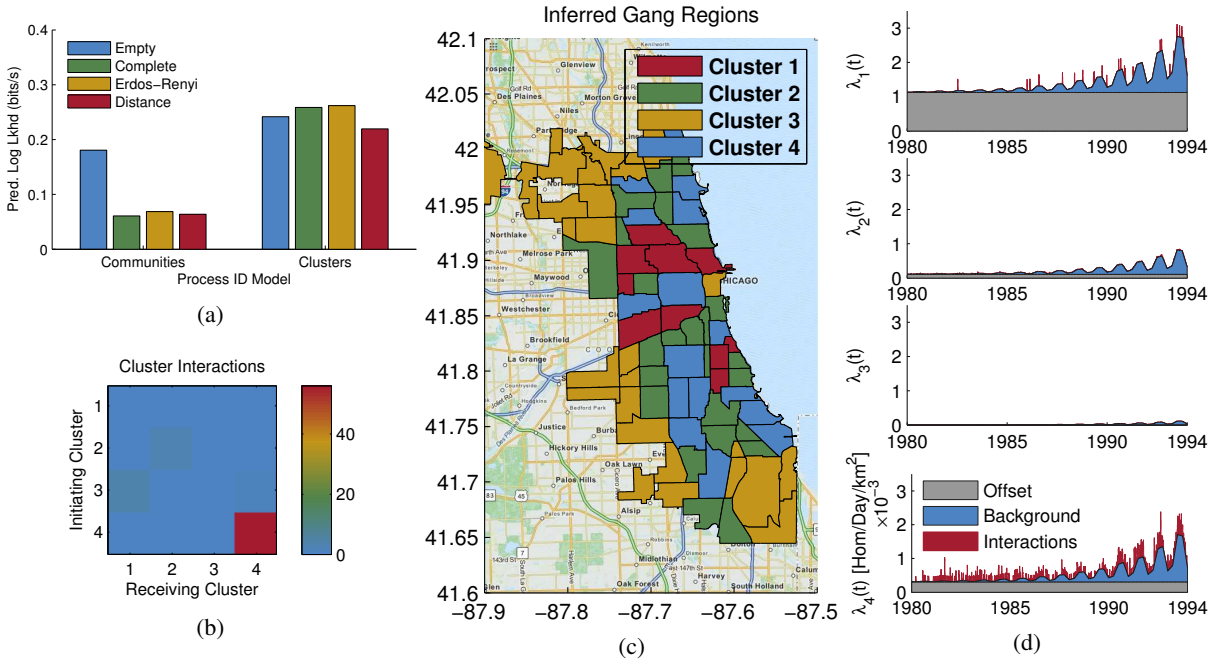


Figure 7: Inferred interactions among clusters of community areas in the city of Chicago. (a) Predictive log likelihood for "communities" and "clusters" process identity models and four graph models. Panels (b-d) present results for the model with the highest predictive log likelihood: an Erdős-Renyi graph with $K = 4$ clusters. (b) The weighted interaction network in units of induced homicides over the training period (1980-1993). (c) Inferred clustering of the 77 community areas. (d) The intensity for each cluster, broken down into the offset, the shared background rate, and the interactions (units of $10^{-3}$ homicides per day per square kilometer).

work Hawkes model with LGCP background rates under two graph priors. The models are trained on four days of data and tested on the fifth. Though the network Hawkes is slightly outperformed by the standard Hawkes, the difference is small relative to the performance improvement from considering interactions, and the inferred network parameters provide interpretable insight into the market structure.

In the latent distance model for $\boldsymbol{A}$, each stock has a latent embedding $\boldsymbol{x}_k \in \mathbb{R}^2$ such that nearby stocks are more likely to interact, as described in Section 2.3. Figure 6 shows a sample from the posterior distribution over embeddings in $\mathbb{R}^2$ for $\rho = 0.2$ and $\tau = 1$. We have plotted stocks in the six largest sectors, as listed on Bloomberg.com. Some sectors, notably energy and financials, tend to cluster together, indicating an increased probability of interaction between stocks in the same sector. Other sectors, such as consumer goods, are broadly distributed, suggesting that these stocks are less influenced by others in their sector. For the consumer industry, which is driven by slowly varying factors like inventory, this may not be surprising.

The Hinton diagram in the bottom panel of Figure 6 shows the top 4 eigenvectors of the interaction network. All eigenvalues are less than 1, indicating that the system is stable. The top row corresponds to first eigenvector ($\lambda_{\text{max}} = 0.74$). Apple (AAPL), J.P. Morgan (JPM), and Exxon Mobil (XOM) have notably large entries in the eigenvector, suggesting that their activity will spawn cascades of self-excitation.

## 7. Gangs of Chicago

In our final example, we study spatiotemporal patterns of gang-related homicide in Chicago. Sociologists have suggested that gang-related homicide is mediated by underlying social networks and occurs in mutually-exciting, retaliatory patterns (Papachristos, 2009). This is consistent with a spatiotemporal Hawkes process in which processes correspond to gang territories and homicides incite further homicides in rival territories.

We study gang-related homicides between 1980 and 1995 (Block et al., 2005). Homicides are labeled by the community in which they occurred. Over this time-frame there were $N = 1637$ gang-related homicides in the 77 communities of Chicago.

We evaluate our model with an event-prediction task, training on 1980-1993 and testing on 1994-1995. We use a LGCP temporal background rate in all model variations. Our baseline is a single process with a uniform spatial rate for the city. We test two process identity models: a) the "community" model, which considers each community a separate process, and b) the "cluster" model, which groups communities into processes. The number of clusters is chosen by cross-validation (see supplementary material). For each process identity model, we compare four graph models: a) independent LGCPs (*empty*), b) a standard Hawkes process with all possible interactions (*complete*), c) a net-

work Hawkes model with a sparsity-inducing Erdős-Renyi graph prior, and d) a network Hawkes model with a latent distance model that prefers short-range interactions.

The community process identity model improves predictive performance by accounting for higher rates in South and West Chicago where gangs are deeply entrenched. Allowing for interactions between community areas, however, results in a decrease in predictive power due to overfitting (there is insufficient data to fit all $77^2$ potential interactions). Interestingly, sparse graph priors do not help. They bias the model toward sparser but stronger interactions which are not supported by the test data. These results are shown in the "communities" group of Figure 7a. Clustering the communities improves predictive performance for all graph models, as seen in the "clusters" group. Moreover, the clustered models benefit from the inclusion of excitatory interactions, with the highest predictive log likelihoods coming from a four-cluster Erdős-Renyi graph model with interactions shown in Figure 7b. Distance-dependent graph priors do not improve predictive performance on this dataset, suggesting that either interactions do not occur over short distances, or that local rivalries are not substantial enough to be discovered in our dataset. More data is necessary to conclusively say which.

Looking into the inferred clusters in Figure 7c and their rates in 7d, we can interpret the clusters as "safe suburbs" in gold, "buffer neighborhoods" in green, and "gang territories" in red and blue. Self-excitation in the blue cluster (Figure 7b) suggests that these regions are prone to bursts of activity, as one might expect during a turf-war. This interpretation is supported by reports of "a burst of street-gang violence in 1990 and 1991" in West Englewood ($41.77°$N, $-87.67°$W) (Block & Block, 1993).

Figure 7d also shows a significant increase in the homicide rate between 1989 and 1995, consistent with reports of escalating gang warfare (Block & Block, 1993). In addition to this long-term trend, homicide rates show a pronounced seasonal effect, peaking in the summer and tapering in the winter. A LGCP with a quadratic kernel point-wise added to a periodic kernel captures both effects.

## 8. Related Work

Gomez-Rodriguez et al. (2010) introduced one of the earliest algorithms for discovering latent networks from cascades of events. They developed a highly scalable approximate inference algorithm, but they did not explore the potential of random network models or emphasize the point process nature of the data. Simma & Jordan (2010) studied this problem from the context of Hawkes processes and developed an expectation-maximization inference algorithm. We have adapted their latent variable formulation in our fully-Bayesian inference algorithm and introduced a framework for prior distributions over the latent network.

Others have considered special cases of the model we have proposed. Blundell et al. (2012) combine Hawkes processes and the Infinite Relational Model (a specific exchangeable graph model with an Aldous-Hoover representation) to cluster processes and discover interactions. Cho et al. (2013) applied Hawkes processes to gang incidents in Los Angeles. They developed a spatial Gaussian mixture model (GMM) for process identities, but did not explore structured network priors. We experimented with this process identity model but found that it suffers in predictive log likelihood tests (see supplementary material).

Recently, Iwata et al. (2013) developed a stochastic EM algorithm for Hawkes processes, leveraging similar conjugacy properties, but without network priors. Zhou et al. (2013) have developed a promising optimization-based approach to discovering low-rank networks in Hawkes processes, similar to some of the network models we explored.

Perry & Wolfe (2013) derived a partial likelihood inference algorithm for Hawkes processes with a similar emphasis on structural patterns in the network of interactions. They provide an estimator capable of discovering homophily and other network effects. Our fully-Bayesian approach generalizes this method to capitalize on recent developments in random network models (Lloyd et al., 2012).

Finally, generalized linear models (GLMs) are widely used in computational neuroscience (Paninski, 2004). GLMs allow for both excitatory and inhibitory interactions, but, as we have shown, when the data consists of purely excitatory interactions, Hawkes processes outperform GLMs in link- and event-prediction tests.

## 9. Conclusion

We developed a framework for discovering latent network structure from spiking data. Our auxiliary variable formulation of the multivariate Hawkes process supported arbitrary Aldous-Hoover graph priors, log Gaussian Cox process background rates, and models of unobserved process identities. Our parallel MCMC algorithm allowed us to reason about uncertainty in the latent network in a fully-Bayesian manner. We leveraged results from random matrix theory to analyze the conditions under which random network models will be stable, and our applications uncovered interpretable latent networks in a variety of synthetic and real-world problems.

**Acknowledgments** We thank Eyal Dechter and Leslie Valiant for their many contributions. S.W.L. was supported by a NDSEG Fellowship. This work was partially funded by DARPA Young Faculty Award N66001-12-1-4219.

## References


Aldous, David J. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

Block, Carolyn R and Block, Richard. *Street gang crime in Chicago*. US Department of Justice, Office of Justice Programs, National Institute of Justice, 1993.

Block, Carolyn R, Block, Richard, and Authority, Illinois Criminal Justice Information. Homicides in Chicago, 1965-1995. ICPSR06399-v5. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], July 2005.

Blundell, Charles, Heller, Katherine, and Beck, Jeffrey. Modelling reciprocating relationships with Hawkes processes. *Advances in Neural Information Processing Systems*, 2012.

Cho, Yoon Sik, Galstyan, Aram, Brantingham, Jeff, and Tita, George. Latent point process models for spatial-temporal networks. *arXiv:1302.2671*, 2013.

Daley, Daryl J and Vere-Jones, David. *An introduction to the theory of point processes.* Springer-Verlag, New York, 1988.

Goldenberg, Anna, Zheng, Alice X, Fienberg, Stephen E, and Airoldi, Edoardo M. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

Gomez-Rodriguez, Manuel, Leskovec, Jure, and Krause, Andreas. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1019–1028. ACM, 2010.

Hawkes, Alan G. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83, 1971.

Hoff, Peter D. Modeling homophily and stochastic equivalence in symmetric relational data. *Advances in Neural Information Processing Systems 20*, 20:1–8, 2008.

Hoover, Douglas N. Relations on probability spaces and arrays of random variables. *Technical report, Institute for Advanced Study, Princeton*, 1979.

Iwata, Tomoharu, Shah, Amar, and Ghahramani, Zoubin. Discovering latent influence in online social activities via shared cascade Poisson processes. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 266–274. ACM, 2013.

Liben-Nowell, David and Kleinberg, Jon. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

Lloyd, James Robert, Orbanz, Peter, Ghahramani, Zoubin, and Roy, Daniel M. Random function priors for exchangeable arrays with applications to graphs and relational data. *Advances in Neural Information Processing Systems*, 2012.

Mohamed, Shakir, Ghahramani, Zoubin, and Heller, Katherine A. Bayesian and L1 approaches for sparse unsupervised learning. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 751–758, 2012.

Møller, Jesper, Syversveen, Anne Randi, and Waagepetersen, Rasmus Plenge. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25 (3):451–482, 1998.

Murray, Iain, Adams, Ryan P., and MacKay, David J.C. Elliptical slice sampling. *Journal of Machine Learning Research: Workshop and Conference Proceedings (AISTATS)*, 9:541–548, 2010.

Paninski, Liam. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262, January 2004.

Papachristos, Andrew V. Murder by structure: Dominance relations and the social structure of gang homicide. *American Journal of Sociology*, 115(1):74–128, 2009.

Perry, Patrick O and Wolfe, Patrick J. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2013.

Silverstein, Jack W. The spectral radii and norms of large dimensional non-central random matrices. *Stochastic Models*, 10(3):525–532, 1994.

Simma, Aleksandr and Jordan, Michael I. Modeling events with cascades of Poisson processes. *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

Zhou, Ke, Zha, Hongyuan, and Song, Le. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 16, 2013.